生物統計在蔬菜育種上之應用

胡凱康

國立台灣大學

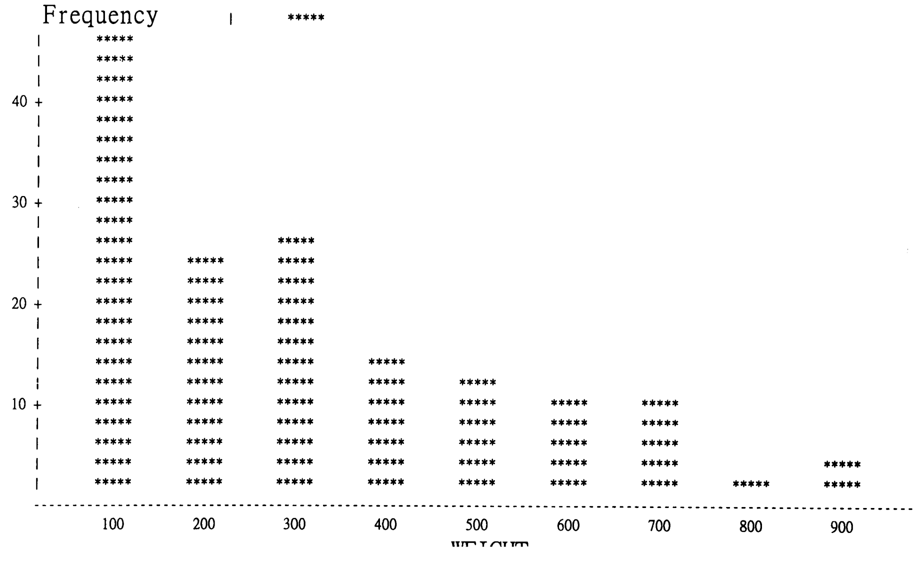
生物的變異性與統計

生物的性狀通常表現出比工業產品高的變異性，這是由於生物有複雜而歧異的遺傳控制，與生長發育過程中對於微小環境差異的反應而產生的偏差。而農業試驗的目的通常在比較不同品種或處理效應的差異，統計就是提供我們如何設計來控制不可預期的環境差異與分析資料時在某種信心水準下獲得結論的方法。我們以下列紅蘿蔔產量的資料來說明生物性狀的變異。

紅蘿蔔重量 (g)

405	109	549	72	32	248	221	234	287	15
388	166	106	257	62	325	673	610	709	494
498	194	499	144	93	203	179	70	717	185
373	380	353	640	509	74	238	149	36	95
529	147	376	662	86	217	380	589	113	496
78	172	318	120	296	1017	197	202	116	129
66	201	292	59	341	606	97	284	915	512
290	87	39	429	643	191	305	258	252	385
216	66	288	670	390	100	579	106	245	426
599	874	597	656	80	175	106	543	340	134
325	55	274	18	71	211	331	254	87	77
749	21	446	88	283	212	31	286	343	91
556	93	132	54	657	223	325	121	436	124
771	505	251	173	258	484	28	813	663	318
393	157	325	87						

上表中之資料可用下列的相對頻度圖表示。下圖顯示出每棵紅蘿蔔重量的分佈，其變異的範圍大約為100—900，而且超過40%的紅蘿蔔其重量在100g左右。

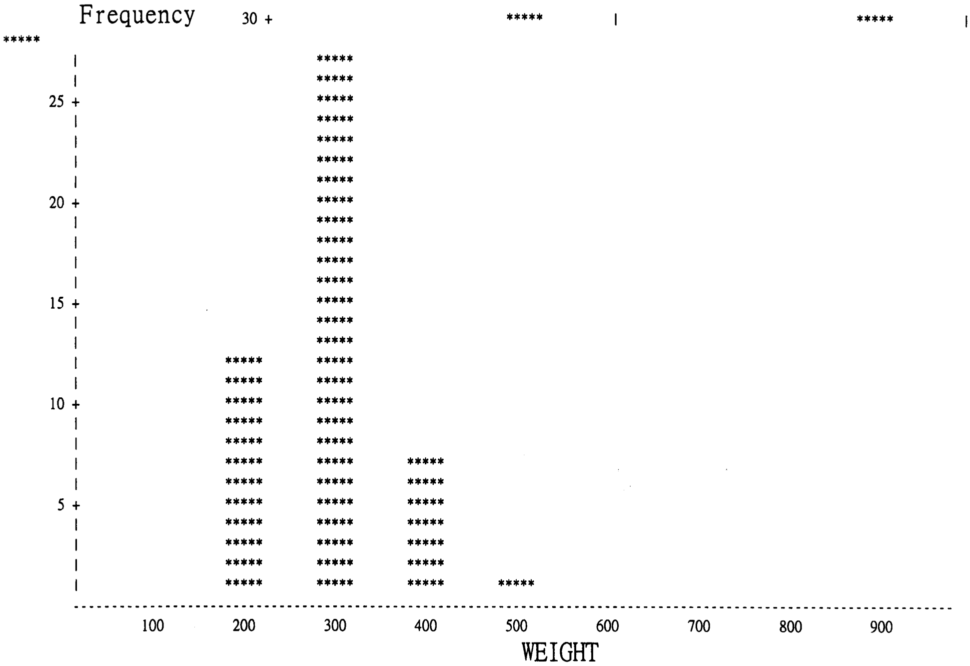


平均＝301.9 標準偏差＝221.3 最大值＝1017 最小值＝15

一般來說，試驗的單位是小區，因此我們想知道每個單位的變異。下表是紅蘿蔔每小區50棵的平均重量 (g)

200.6	377.8	276.3	301.8	291.6	400.2
296.8	207.3	273.4	371.4	228.2	319.5
266.4	263.8	359.9	279.3	318.7	242.3
287.9	261.8	301.1	174.6	228.1	335.3
243.5	308.2	494.9	275.7	303.8	237.4
250.8	276.0	298.8	246.8	266.6	262.4
378.9	306.4	224.9	315.3	237.3	274.9
339.2	236.9	378.9	436.3	346.4	268.6
259.5	291.0				

上表中之資料可用相對頻度圖表示。下圖顯示其變異範圍大致為200—500，而且大約集中在平均值300g附近。



平均值＝292.5 標準偏差＝61.2 最大值＝494.9 最小值＝174.6

樣品與特定值的比較（*t*-test）

假設我們要求每一個小麥的新品種都必須具有每英畝兩噸的產量，某一個新品種在單一試驗時得到2.5噸/英畝的產量，但是單一試驗的結果並不能讓我們對所要回答的問題得到結果，因為產量本身也是一個逢機變數，這一次的結果並不能代表下一次試驗的結果也會比標準的兩噸/英畝高。多做五次試驗，得到2.1, 2.4, 2.0, 2.6 與2.2噸/英畝的結果。以六次試驗的結果求變方的估計值，均方 $s^2 = \sum(x-\bar{x})^2/(n-1) = 0.056$,因此標準偏差 $s = 0.237$，平均值的標準偏差為 $0.237/\sqrt{6} = 0.097$，而可得到 $t = (2.3-2)/0.097 = 3.1$。查表得到 $t_{\alpha=0.05,df=5} = \pm 2.57$，因為實測 t 值落於±2.57的範圍之外，因此我們有95％的把握新品種的產量確實比標準值高。

兩樣品間的比較

因為每一季的生長條件不同，在測驗新品種時，我們通常不採用以上的方法，而是以一個標準品種與新品種一起種植，如：

新品種： 2.5 2.1 2.4 2.0 2.6 2.2

標準品種： 2.2 1.9 1.8 2.1 2.1 1.7 2.3 2.0 1.7 2.2

$\bar{x}_1 = 2.3$，$s_1^2 = 0.056$，

$\bar{x}_2 = 2.0$，$s_2^2 = 0.047$，

合併的均方為 $s^2 = \dfrac{5(0.056)+9(0.047)}{6+10-2} = 0.050$,

兩樣品差值的標準偏差 $= \sqrt{s^2(\dfrac{1}{n_1}+\dfrac{1}{n_2})} = \sqrt{0.05\times(\dfrac{1}{6}+\dfrac{1}{10})} = 0.1155$,

$t = \dfrac{2.3-2.0}{0.1155} = 2.60$, 查 t 分佈表，自由度為 $n_1+n_2-2=14, t_{\alpha=0.05,df=14} = 2.14$

因此我們有95%的把握，新品種比標準品種的產量高。

簡單的SAS程式及輸出如下：

```
data wheat;
  input var $ yield @@;
  cards;
n 2.5 n 2.1 n 2.4 n 2.0 n 2.6 n 2.2
s 2.2 s 1.9 s 1.8 s 2.1 s 2.1 s 1.7 s 2.3 s 2.0 s 1.7 s 2.2
;
proc ttest data=wheat;
  class var;
  var yield;
run;
```

```
VAR      N     Mean       Std Dev        Std Error      Minimum     Maximum
---------------------------------------------------------------------------
n        6     2.30     0.23664319     0.09660918      2.0000      2.6000
s       10     2.00     0.21602469     0.06831301      1.7000      2.3000

Variances          T        DF     Prob>|T|
----------------------------------------
Unequal       2.5355       9.9      0.0299
Equal         2.5981      14.0      0.0211

For H0: Variances are equal, F' = 1.20    DF = (5,9)    Prob>F' = 0.7634
```

報表顯示這兩個品種有相同之變異（P=0.76），因此我們將兩個均方（0.056及0.047）合併成為共同均方0.050。

兩個處理的完全逢機設計
（Completely Randomized Design, CR Design）

設計的目的在於控制執行試驗的環境所產生的變異，完全逢機試驗設計適用於田區的地力變異為逢機分佈的情況，它的一個特點是各重複的數目並不一定相等，甚至對於已知變異較高的處理，可以採用較多的重複次數來控制其機差。

A 24	B 28	B 31	B 26
A 27	B 23	A 21	B 29
A 22	A 20		

以 *t*-test 的方法計算：

$s_A^2 = 7.7$，$\bar{x}_A = 22.8$，

$s_B^2 = 9.3$，$\bar{x}_B = 27.4$，

$s^2 = (4 \times 7.7 + 4 \times 9.3)/(5+5-2) = 8.5,$

兩處理間差異的標準偏差為 $\sqrt{8.5 \times \left(\frac{1}{5} + \frac{1}{5}\right)} = 1.84,$

$t = \frac{27.4 - 22.8}{1.84} = 2.5$，

$t_{\alpha=0.05, df=8} = 2.31,$ 所以A與B品種間有顯著的差異。

用變方分析法計算：

```
data a;
  input variety $ yield @@;
  cards;
A 24 B 28 B 31 B 26
A 27 B 23 A 21 B 29
A 22 A 20
;
proc anova data=a;
  class variety;
  model yield=variety;
run;
```

```
Analysis of Variance Procedure
Dependent Variable: YIELD

Source          DF     Sum of Squares     Mean Square      F Value     Pr > F
Model            1       52.90000000      52.90000000        6.22      0.0372
Error            8       68.00000000       8.50000000
Corrected Total  9      120.90000000

R-Square               C.V.          Root MSE        YIELD Mean
0.437552             11.61544       2.91547595       25.10000000

Source          DF        Anova SS         Mean Square      F Value     Pr > F
VARIETY          1       52.90000000      52.90000000        6.22       0.0372
```

只有兩個處理時，*t*-test與變方分析的結果是完全一樣的。

多個處理項目的CR設計

品種	A	B	C	D
產量	25.12	40.25	18.30	28.55
	17.25	35.25	22.60	28.05
	26.42	31.98	25.90	33.20
	16.08	36.52	15.05	31.68
	22.15	43.32	11.42	30.32
	15.92	37.10	23.68	27.58
平均	20.49	37.40	19.49	29.90
均方	22.04	15.61	30.91	4.97

首先要回答的問題是「品種之間有沒有差異？」如果有，其次要回答的問題是「哪些品種之間有差異？」

```
data a;
  input variety $ yield @@;
  cards;
A 25.12 B 40.25 C 18.30 D 28.55
A 17.25 B 35.25 C 22.60 D 28.05
A 26.42 B 31.98 C 25.90 D 33.20
A 16.08 B 36.52 C 15.05 D 31.68
A 22.15 B 43.32 C 11.42 D 30.32
A 15.92 B 37.10 C 23.68 D 27.58
;
proc anova data=a;
  class variety;
  model yield=variety;
run;
```

```
Analysis of Variance Procedure

Dependent Variable: YIELD
                                  Sum of            Mean
Source                  DF       Squares           Square     F Value     Pr > F

Model                    3     1291.477146      430.492382     23.42      0.0001
```

```
Error                  20      367.653150      18.382658

Corrected Total        23     1659.130296

              R-Square          C.V.        Root MSE            YIELD Mean

              0.778406      15.98596        4.287500              26.82042


Source                 DF       Anova SS   Mean Square   F Value      Pr > F

VARIETY                 3    1291.477146    430.492382     23.42      0.0001
```

若想要知道是哪些品種之間有差異，可用的方法很多，其中一種為鄧肯氏多重變域法。修改的SAS程式及其結果如下：

```
proc anova;
  class variety;
  model yield=variety;
  means variety /Duncan;
run;
```

```
Duncan's Multiple Range Test for variable: YIELD

NOTE: This test controls the type I comparisonwise error rate, not the
      experimentwise error rate

Alpha= 0.05  df= 20  MSE= 18.38266

Number of Means      2     3     4
Critical Range   5.164 5.420 5.583

Means with the same letter are not significantly different.

Duncan Grouping             Mean      N  VARIETY

                 A           37.403      6  B

                 B           29.897      6  D

                 C           20.490      6  A
                 C
                 C           19.492      6  C
```

其他常用的方法還有LSD和Sheffe的方法。程式的修改與上同，只要將比較方法的keyword填入Duncan的位置即可，其結果分列如下：

```
T tests (LSD) for variable: YIELD

NOTE: This test controls the type I comparisonwise error rate not the
      experimentwise error rate.
Alpha= 0.05  df= 20  MSE= 18.38266
```

```
Critical Value of T= 2.09
Least Significant Difference= 5.1636
Means with the same letter are not significantly different.

T Grouping                Mean       N  VARIETY
                A           37.403       6  B
                B           29.897       6  D
                C           20.490       6  A
                C
                C           19.492       6  C
```

```
Scheffe's test for variable: YIELD

NOTE: This test controls the type I experimentwise error rate but generally
      has a higher type II error rate than REGWF for all pairwise comparisons

Alpha= 0.05  df= 20  MSE= 18.38266
Critical Value of F= 3.09839
Minimum Significant Difference= 7.547

Means with the same letter are not significantly different.

Scheffe Grouping              Mean       N  VARIETY

              A            37.403      6   B
              A
              A            29.897      6   D

              B            20.490      6   A
              B
              B            19.492      6   C
```

註：

Type I error: reject H_0 | H_0 is true

Type II error: fail to reject H_0 | H_0 is false

逢機完全區集設計
(Randomized Complete Block Design, RCB Design)

當試驗環境的變異有趨勢可循，例如有坡度的試驗田，溫室中微氣象的變異，或在不同時間的接種抗病評估試驗等，都可以視為區集。簡而言之，區集的設立是當區集內的變異小於區集間的變異時，一種控制試驗機差很有效的手段。

I		II		III	
A	B	D	A	C	A
C	D	B	C	B	D

下例為測試七種土壤種類及栽培介質對金魚草莖長的作用，每一個區集中均包含七種土壤的種類，

```
data plants;
  input type $ @;
  do block=1 to 3;
    input stemleng @;
    output;
  end;
  cards;
CLARION 32.7 32.3 31.5
CLINTON 32.1 29.7 29.1
KNOX    35.7 35.9 33.1
O'NWILL 36.0 34.2 31.2
COMPOST 31.8 28.0 29.2
WABASH  38.2 37.8 31.9
WEBSTER 32.5 31.1 29.8
;
proc anova data=plants;
  class type block;
  model stemleng=block type;
run;
```

```
Analysis of Variance Procedure

Dependent Variable: STEMLENG
                                    Sum of          Mean
Source                  DF         Squares        Square    F Value      Pr >F

Model                    8     141.5523810    17.6940476      10.71     0.0002

Error                   12      19.8171429     1.6514286

Corrected Total         20     161.3695238

                  R-Square            C.V.      Root MSE         STEMLENG Mean

                  0.877194        3.946573      1.285079              32.56190


Source                  DF        Anova SS   Mean Square    F Value      Pr > F

BLOCK                    2      38.6895238    19.3447619      11.71      0.0015
TYPE                     6     102.8628571    17.1438095      10.38      0.0004
```

逢機完全區集設計的數學模式為： $y_{ij}=\mu+b_i+t_j+e_{ij}$

其變方分析表中各變因的計算方式如下：

變異來源	s.s.	d.f.	m.s.
區集	$\frac{B_1^2+B_2^2+\cdots+B_b^2}{t}-\frac{[\sum(x)]^2}{t}$	b-1	
處理	$\frac{T_1^2+T_2^2+\cdots T_t^2}{b}-\frac{[\sum(x)]^2}{n}$	t-1	
機差	減後的剩餘	(b-1)(t-1)	s^2
總計	$\sum(x^2)-\frac{[\sum(x)]^2}{n}$	n-1	

可以看出，總計平方和與處理平方和的計算方法是一樣的，只要區集之間有明顯的差異（區集平方和大），機差平方和就會變小，因而達到控制機差的目標。

複因子試驗設計

一般說來，通常一個試驗中都牽涉一個以上的因子，今以兩因子的試驗為例，其設計是採用CR設計，所得的資料其數學模式為 $y_{ijk}=\mu+\tau_i+\beta_j+(\tau\beta)_{ij}+\varepsilon_{ijk}$，是中 $i=1,\ldots,a$ 表示A因子有a個變級， $j=1,\ldots,b$ 表示B因子有b個變級， τ_i ， β_j 與 $(\tau\beta)_{ij}$ 分別表示A因子的主效應，B因子的主效應與AB之交感效應。

當兩個主效應都是固定型效應時，複因子試驗的均方期望值如下：

變因	均方期望值
τ_i	$\sigma^2+\frac{bn\sum\tau_i^2}{a-1}$
β_j	$\sigma^2+\frac{an\sum\beta_j^2}{b-1}$
$(\tau\beta)_{ij}$	$\sigma^2+\frac{n\sum\sum(\tau\beta)_{ij}^2}{(a-1)(b-1)}$
ε_{ijk}	σ^2

今以一實例說明兩因子試驗用RCB設計的分析方法。

種子批號	播種量（磅／畝）	區集 I	區集 II	區集 III
T	1.5	4.20	4.94	4.45
	2	4.36	3.50	4.17
	2.5	5.40	4.55	5.75
	3	5.15	4.40	3.90
H	1.5	2.82	3.14	3.80
	2	3.74	4.43	2.92
	2.5	4.82	3.90	4.50
	3	4.57	5.32	4.35

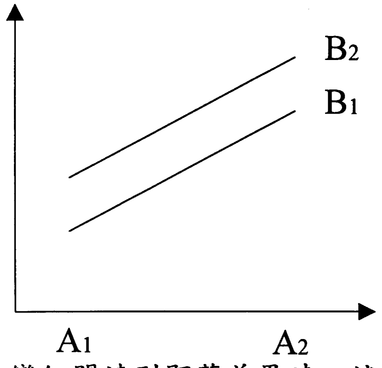
```
data carrot;
  input batch $ @;
  do j=1 to 4;
    input rate @;
    do block=1 to 3;
      input yield @;
      output;
    end;
  end;
  cards;
T   1.5      4.20      4.94      4.45
      2      4.36      3.50      4.17
    2.5      5.40      4.55      5.75
      3      5.15      4.40      3.90
H   1.5      2.82      3.14      3.80
      2      3.74      4.43      2.92
    2.5      4.82      3.90      4.50
      3      4.57      5.32      4.35
;
proc glm;
  class block batch rate;
  model yield=block batch rate batch*rate;
  means rate /LSD;
run;
```

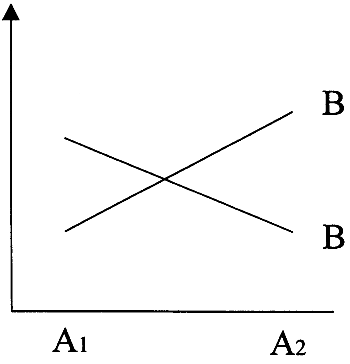
```
Dependent Variable: YIELD
                                    Sum of          Mean
Source                    DF       Squares        Square    F Value      Pr > F

Model                      9    8.23490000    0.91498889       2.70      0.0466

Error                     14    4.74170000    0.33869286

Corrected Total           23   12.97660000

                    R-Square          C.V.      Root MSE             YIELD Mean

                    0.634596      13.55002      0.581973               4.295000


Source                    DF     Type I SS   Mean Square    F Value      Pr > F

BLOCK                      2    0.09910000    0.04955000       0.15      0.8652
BATCH                      1    1.73881667    1.73881667       5.13      0.0398
RATE                       3    4.41463333    1.47154444       4.34      0.0232
BATCH*RATE                 3    1.98235000    0.66078333       1.95      0.1678

          T tests (LSD) for variable: YIELD

          NOTE: This test controls the type I comparisonwise error rate not
                the experimentwise error rate.

                    Alpha= 0.05  df= 14  MSE= 0.338693
```

```
                    Critical Value of T= 2.14
               Least Significant Difference= 0.7207

     Means with the same letter are not significantly different.

          T Grouping            Mean      N  RATE

                   A          4.8200      6  2.5
                   A
                   A          4.6150      6  3

                   B          3.8917      6  1.5
                   B
                   B          3.8533      6  2
```

複因子試驗設計中，最需要注意的問題之一就是交感項檢定後對於主效應的解釋問題。簡單來說，當交感項顯著時，試驗者不可以針對某一主效應的變級間進行比較而不談另一個主效應。

左圖中顯示的是沒有交感的情況，這時可以單獨比較主效應中變級的優劣，例如在變級間達到顯著差異時，試驗者可以說B_2品種比B_1品種的產量高，或A_2環境的生產力較A_1為高。右圖顯示兩個主效應之間有交感的存在，此時試驗者無法只說哪一個品種的產量較高，而必須指明在A_2環境中，B_1品種的產量比B_2品種高。





雖然在品系比較試驗中的主效應都是固定型效應，但複因子試驗中，也有些時候主效應都是逢機型效應或是兩者都有的混合型模式。分析時，固定型、逢機型與混合型模式之間最大的不同在於各變因的檢定項不同。

變因	逢機型均方期望值	混合型均方期望值
τ_i	$\sigma^2+n\sigma_{\tau\beta}^2+bn\sigma_\tau^2$	$\sigma^2+n\sigma_{\tau\beta}^2+\frac{bn\sum\tau_i^2}{a-1}$
β_j	$\sigma^2+n\sigma_{\tau\beta}^2+an\sigma_\beta^2$	$\sigma^2+an\sigma_\beta^2$
$(\tau\beta)_{ij}$	$\sigma^2+n\sigma_{\tau\beta}^2$	$\sigma^2+n\sigma_{\tau\beta}^2$
ε_{ijk}	σ^2	σ^2

複因子試驗設計與單因子的試驗設計來比較，除了可以得到單因子試驗中無法觀測的交感效應之外，另一個優點是它有隱藏重複，試驗的效率較單因子試驗高。

要注意的是，在本節中所舉的數例只是為了舉例方便，其實就該例而言，所列的分析方法並不是最恰當的。在上例中，種子批號是屬於「質」的因子，但播種量的各變級是屬於「量」的因子。以上的分析方法忽略了播種量的變級間有漸進的程度上的差異，而把它們當作完全無關的變級來處理。要觀測「量」的因子各變級的效應，可將該因子及其有關的交感效應作單自由度的劃分，以檢定該因子的一次、二次及三次曲線的效應。

裂 區 設 計

裂區設計是複因子試驗設計的一種變形，通常用於當其中一因子的施用較為困難時，例如有多個品系參加的灌溉或肥料試驗，因為灌溉或肥料的施用很容易影響到鄰近的小區，因此將相同處理的小區集合起來形成一個大區，例如在下圖中的試驗包含四個重複，每重複有三個大區，每個大區則有六個小區，x, y, z分別為大區的三個處理，A, B, C, D, E, F則分別為小區的六個處理。

	I			II	III	IV
x	A	B	D			
	F	E	C			
y	C	E	F			
	B	D	A			
z	C	D	B			
	F	A	E			

裂區設計的數學模式為

$$y_{ijk} = \mu + \rho_i + \alpha_j + (\rho\alpha)_{ij} + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk}$$

當主區與副區（裂區）都是固定型效應時，裂區試驗的均方期望值如下：

		變因	均方期望值
主區	重複	ρ_i	$\sigma_\varepsilon^2 + bc\sigma_\rho^2$
	主效應	α_j	$\sigma_\varepsilon^2 + c\sigma_\delta^2 + \dfrac{ac\sum\alpha_j^2}{b-1}$
	機差	$(\rho\alpha)_{ij}$	$\sigma_\varepsilon^2 + c\sigma_\delta^2$
副區	主效應	β_k	$\sigma_\varepsilon^2 + \dfrac{ab\sum\beta_k^2}{c-1}$
	交感	$(\alpha\beta)_{jk}$	$\sigma_\varepsilon^2 + \dfrac{a\sum\sum(\alpha\beta)_{jk}^2}{(b-1)(c-1)}$
	機差	ε_{ijk}	σ_ε^2

下例為種在框架內的六個萵苣品種對三個不同綠化日期的反應，

綠化日期	品種	I	II	III	IV
x	A	11.8	7.5	9.7	6.4
	B	8.3	8.4	11.8	8.5
	C	9.2	10.6	11.4	7.2
	D	15.6	10.8	10.3	14.7
	E	16.2	11.2	14.0	11.5
	F	9.9	10.8	4.8	9.8
y	A	9.7	8.8	12.5	9.4
	B	5.4	12.9	11.2	7.8
	C	12.1	15.7	7.6	9.4
	D	13.2	11.3	11.0	10.7
	E	16.5	11.1	10.8	8.5
	F	12.5	14.3	15.9	7.5
z	A	7.0	9.1	7.1	6.3
	B	5.7	8.4	6.1	8.8
	C	3.3	6.9	1.0	2.6
	D	12.6	15.4	14.2	11.3
	E	12.6	12.3	14.4	14.1
	F	10.2	11.6	10.4	12.2

```
data lettuce;
  input date $ @;
  do j=1 to 6;
    input var $ @;
    do block=1 to 4;
      input yield @;
      output;
    end;
  end;
  cards;
x   A   11.8   7.5   9.7   6.4
    B    8.3   8.4  11.8   8.5
    C    9.2  10.6  11.4   7.2
    D   15.6  10.8  10.3  14.7
    E   16.2  11.2  14.0  11.5
    F    9.9  10.8   4.8   9.8
y   A    9.7   8.8  12.5   9.4
    B    5.4  12.9  11.2   7.8
    C   12.1  15.7   7.6   9.4
    D   13.2  11.3  11.0  10.7
    E   16.5  11.1  10.8   8.5
    F   12.5  14.3  15.9   7.5
z   A    7.0   9.1   7.1   6.3
    B    5.7   8.4   6.1   8.8
    C    3.3   6.9   1.0   2.6
    D   12.6  15.4  14.2  11.3
    E   12.6  12.3  14.4  14.1
    F   10.2  11.6  10.4  12.2
;
proc glm;
  class block date var;
  model yield=block date block*date var date*var;
  test h=date e=block*date;
run;
```

注意在model的陳述中，block*var與block*date*var都沒有出現，因此這兩項在計算的過程中會被合併起來，並作為全部試驗因子的檢定項。針對主區的效應，必須另以test陳述指定所要用的檢定項。

```
General Linear Models Procedure

Dependent Variable: YIELD
                                  Sum of          Mean
Source                  DF       Squares        Square    F Value      Pr > F

Model                   26   535.1183333    20.5814744       4.08      0.0001

Error                   45   227.2766667     5.0505926

Corrected Total         71   762.3950000

                  R-Square          C.V.      Root MSE            YIELD Mean

                  0.701891      21.87204      2.247352              10.27500

Source                  DF     Type I SS   Mean Square    F Value      Pr > F

BLOCK                    3    29.3427778     9.7809259       1.94      0.1372
DATE                     2    38.0033333    19.0016667       3.76      0.0308
BLOCK*DATE               6    43.5655556     7.2609259       1.44      0.2216
VAR                      5   260.5083333    52.1016667      10.32      0.0001
DATE*VAR                10   163.6983333    16.3698333       3.24      0.0032


Tests of Hypotheses using the Type III MS for BLOCK*DATE as an error term

Source                  DF   Type III SS   Mean Square    F Value      Pr > F

DATE                     2   38.00333333   19.00166667       2.62      0.1524
```

將上列報表整理為常見的格式如下：

變異來源	自由度	均方	F值
區集	3	9.78	
綠化日期（主區）	2	19.00	$2.62^{n.s.}$
主區機差	6	7.26	
品種（副區）	5	52.10	10.32**
綠化日期×品種	10	16.37	3.24**
副區機差	45	5.05	

以上的結果顯示副區的機差比主區的小，這是可預期的結果，因為同一個主區中的各副區的反應較為近似。我們可以說，一般的複因子試驗中對所有的處理組合都有一致的準確度，相對而言，裂區設計犧牲了一部份主區間比較的準確度，但提高了副區間比較的準確度，因此在試驗條件許可下，通常建議將比較重要的因子作為副區。

變積分析

有些時候即使我們盡量控制試驗中的變異性，仍然在小區間出現有相當程度的變異，若這些變異是可以測量的連續性變異，可用變積分析來加以校正，達到控制機差的目的。

在下例中，試驗者想要知道蘋果園中不同覆蓋作物對蘋果產量的影響。處理共有六種（A至F），試驗分為四個區集，y為連續四年的蘋果產量（以英磅計算）。第一部份的分析依照一般單一因子RCB設計。

區集	I		II		III		IV	
處理	x	y	x	y	x	y	x	y
A	8.2	287	9.4	290	7.7	254	8.5	307
B	8.2	271	6.0	209	9.1	243	10.1	348
C	6.8	234	7.0	210	9.7	286	9.9	371
D	5.7	189	5.5	205	10.2	312	10.3	375
E	6.1	210	7.0	276	8.7	279	8.1	344
F	7.6	222	10.1	301	9.0	238	10.5	357

```
data fruit;
  input treat $ @;
  do block=1 to 4;
    input x y @;
    output;
  end;
  cards;
A  8.2  287  9.4  290  7.7  254  8.5  307
B  8.2  271  6.0  209  9.1  243 10.1  348
C  6.8  234  7.0  210  9.7  286  9.9  371
D  5.7  189  5.5  205 10.2  312 10.3  375
E  6.1  210  7.0  276  8.7  279  8.1  344
F  7.6  222 10.1  301  9.0  238 10.5  357
;
proc glm;
  class block treat;
  model y=block treat;
run;
```

```
General Linear Models Procedure
Dependent Variable: Y
                              Sum of          Mean
Source              DF       Squares        Square    F Value     Pr >F
Model                8   48602.33333    6075.29167       3.89    0.0113

Error               15   23432.16667    1562.14444

Corrected Total     23   72034.50000

               R-Square          C.V.      Root MSE              Y Mean

               0.674709      14.33326      39.52397            275.7500
```

Source	DF	Type I SS	Mean Square	F Value	Pr > F
BLOCK	3	47852.83333	15950.94444	10.21	0.0006
TREAT	5	749.50000	149.90000	0.10	0.9915

Source	DF	Type III SS	Mean Square	F Value	Pr > F
BLOCK	3	47852.83333	15950.94444	10.21	0.0006
TREAT	5	749.50000	149.90000	0.10	0.9915

結果顯示出比處理均方（149.9）高出十倍以上的機差均方（1562.1）。檢討的結果發現同一區集內的各株果樹之間本來就有很大的變異，為了校正這個因素，便採用四年前的產量記錄（單位為bushel，約36公升）進行變積分析。當時果園中並沒有施加任何額外的處理，因此這個變數，x，代表的是各株本身的生產潛能。

```
proc glm;
  class block treat;
  model y=block treat x;
  lsmeans treat /pdiff;
run;
```

General Linear Models Procedure

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr >F
Model	9	68149.29602	7572.14400	27.29	0.0001
Error	14	3885.20398	277.51457		
Corrected Total	23	72034.50000			

R-Square	C.V.	Root MSE	Y Mean
0.946065	6.041258	16.65877	275.7500

Source	DF	Type I SS	Mean Square	F Value	Pr > F
BLOCK	3	47852.83333	15950.94444	57.48	0.0001
TREAT	5	749.50000	149.90000	0.54	0.7430
X	1	19546.96269	19546.96269	70.44	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
BLOCK	3	13258.06114	4419.35371	15.92	0.0001
TREAT	5	4352.89155	870.57831	3.14	0.0417
X	1	19546.96269	19546.96269	70.44	0.0001

這個時候，我們要看的是Type III平方和，因為Type III平方和為主效應經過校正後的結果，$SS(\tau \mid \mu,\beta,\gamma)$，其中第二項為重複的效應，第三項為連續校正項的效應。經過變積分析（第二部份的程式與分析結果），處理之間是有顯著的差異，並且試驗機差也有了相當好的控制。最後我們若想知道哪些處理間有差異，可用LSMEAN指令要求計算出已經校正後的平均值，並以PDIFF選項要求列出各處理間配對比較時的P值。

```
Least Squares Means

TREAT            Y     Pr > |T| H0: LSMEAN(i)=LSMEAN(j)
            LSMEAN     i/j     1        2        3        4        5        6

A       280.476530     1       .   0.2575   0.5950   0.9566   0.1169   0.0307
B       266.566627     2  0.2575        .   0.5346   0.2398   0.0134   0.2327
C       274.066627     3  0.5950   0.5346        .   0.5608   0.0442   0.0838
D       281.137036     4  0.9566   0.2398   0.5608            0.1181   0.0338
E       300.917469     5  0.1169   0.0134   0.0442   0.1181        .   0.0022
F       251.335712     6  0.0307   0.2327   0.0838   0.0338   0.0022        .

NOTE: To ensure overall protection level, only probabilities associated with
      pre-planned comparisons should be used.
```

參考文獻

Freund, R. J., Littell, R. C. and Spector, P. C. 1986 SAS system for linear models. SAS Institute Inc. Cary, NC.

Mead, R., Curnow, R. N. and Hasted, A. M. 1993 Statistical methods in agricultural and experimental biology. (2nd Ed.) Chapman & Hall. London.

Montgomery, D. C. 1991 Design and analysis of experiments. (3rd Ed.) John Wiley and Sons. New York.

SAS Institute Inc. SAS/Stat User's guide, Release 6.03 Edition. Cary, NC:SAS Institute Inc., 1988. 1028 pp.